Draft of ASA paper
20 October 2004

Background

Form EIA-826 collects information from regulated and unregulated companies that sell or deliver electric power to end users, including electric utilities, energy service providers, and distribution companies. The Form EIA-826 is a monthly survey that prior to 2004 collected state-level sales volumes, sales revenues, and number of customers by end-use sector (residential, commercial, industrial, other (including public street and highway lighting), and total).

The Form EIA-826 uses three Schedules to collect information: Schedule A collects from full service providers (bundled electricity and delivery service to end users); Schedule B collects from marketers that provide electricity only service to end users (without delivery service); and Schedule C collects from utilities that own distribution lines that provide delivery only service to end users.

The respondent list for the EIA-826 consists of the following groups:

Respondent Classifications

Schedule	Respondent Group
Schedule A Electricity Generators	Census of IOU’s
	Sample of non-IOUs
Schedule B Transmission	Census of Wholesalers
Schedule C Distribution	Census of Utilities

Form EIA-861 is used to collect retail sales of electricity and associated revenue by sector from all electric utilities, electricity service providers and distribution companies in the United States on an annual basis. It provides the frame for the monthly EIA-826. Hence the respondents to the EIA-826 are a subset of the respondents to the EIA-861.

Sample Design and Estimation for Non-sampled Companies

Schedule A is completed by a combination of a cut-off sample of full service providers and a census of Investor Owned Utilities (IOUs). Schedule B is completed by a census of marketers, and Schedule C is completed by a census of Utilities that provide delivery only service.

The number of companies reporting on the EIA-861 was 3,214 in 2002 and 3,215 in 2003. The companies reporting on Schedule A of the EIA-826 included 259 IOUs and 149 sampled units in 2002, and 261 IOUs and 170 sampled units in 2003.

The cut-off sample (i.e., sample of largest units) was selected using annual data for two different years to demonstrate that relative standard errors (RSE's) were smaller than 1% for residential, commercial, and industrial revenues, sales, and prices. Initially (in the late 80's and early 90's) estimates were done by State and virtually all units in the cut-off sample were used to estimate for the nonsampled units. Since that time, with changes in the industry and the addition of IOUs and power marketers as certainty units, the cut-off sample was adjusted to maintain a total number on the respondent list of fewer than 450.

Instead of making estimates separately by State, estimates are now made within 11 estimation regions that have similar weather and economic conditions. The 11 estimation regions are Alaska, Hawaii, NEA (CT,DE,DC,ME,MD,MA,NH,NJ,PA,RI,NY,VT); NEC (IA,MI,MN,WI); CEN (IL,IN,KY,MO,OH,TN,WV), NWC (MT,NE,ND,SD,WY); WES (CA,NV); NEW (OR,WA,ID); SEA (AL, FL,GA,NC,SC,VA); SOU (AR,KS,LA,MS,OK,TX); and SWE (AZ,CO,MN,UT).

By region, the sample coverage rate (data reported as a percent of monthly estimated total) is presented for each region in the table below.

AK	88.65%
CEN	78.68%
NEA	95.21%
NEC	83.12%
NWC	79.51%
NWE	75.23%
SEA	76.66%
SOU	68.17%
SWE	86.72%
WES	92.21%

Based on the data published for August 2003, the sample coverage rate by sector is 81% for residential revenue, 79% for residential sales, 88% for commercial revenue, 86% for commercial sales, 84% for industrial revenue and sales, 78% for other revenue, 76% for other sales, 84% for total revenue, and 82% for total sales.

Estimation for nonsampled companies in an estimation group is done using a regression equation of the form

$$y_{is} = \beta_s x_{is} + \varepsilon_{is} \qquad (1)$$

Here y_{is} is the current EIA-826 data for company (i) in estimation region (s), x_{is} is the past EIA-861 data for company (i) in region (s), and ε_{is} is the error term, assumed to be normally distributed with mean 0 and variance $\sigma_s{}^2 x_{is}{}^{2\gamma}$. Based on comparisons that were conducted during the late 1980's and early 1990's, γ is currently taken to be 0.8. The coefficient β_s represents the seasonal or business cycle change from the annual data to the current monthly data in estimation region s. The seasonal or business patterns estimated by $\hat{\beta}_s$ are assumed to apply to the nonsampled companies, as well as the sampled ones. The estimated regression coefficient $\hat{\beta}_s$ is used to predict the monthly data for nonsampled companies, based on their EIA-861 data. Hence, $\hat{y}_{ks} = \hat{\beta}_s x_{ks}$ for the kth nonsampled company in region s. Once the estimated values are available for the nonsampled companies, the estimates and reported values together are used to prepare aggregates for States and other regions.

In 2003 most of the sampling error was associated with the "Other" category. This is due to the fact that the emphasis was on the three main sectors: residential, commercial, and industrial. The RSE tables in the *Electric Power Monthly* (EPM) publication provide information on sampling errors. The total numbers of RSEs greater than 10 in the EPM tables for November 2003 were:

Category	**Revenue ($-Mil)**	**Sales (MWh)**	**Price (¢/KWh)**
Total data elements	300	300	300
Residential RSE >10	1	2	0
Commercial RSE > 10	1	2	0
Industrial RSE > 10	6	5	3
Other RSE >10	19	23	15
Total RSE > 10	1	2	0
Total	28	34	18

The RSEs appear to be relatively stable over time. While the RSE for "Other" tends to be large, it is no longer collected beginning in January 2004. (It has been replaced by a new category, "transportation.")

Any company that appears to have valid monthly data, but whose annual (EIA-861) data is inconsistent with the monthly data is treated as an "additive outlier." An additive outlier company's data are used to form monthly totals but are not used in estimation for other companies.

Evaluation

We have used scatterplots, standardized residual plots and other diagnostic tools in a thorough exploratory analysis to assess the quality of the fit of the model in (1). In this exercise we have used both ten geographical regions (those described above except for HI), and a smaller set of four geographical regions North East (NEA, CEN, and NEC); North West (AK, NEW, NWC); South East (SEA, SOU); and South West (SWE, WES, HI). This diagnostic evaluation has included regressions by region with all data included and with outliers and influential observations removed. We have declared an observation to be an outlier if the absolute value of the standardized residual exceeds 3.5, and deemed an observation to be influential if DFFITS exceeds $2/n^{1/2}$.

The attached scatterplots, standardized residual plots and summary statistics are representative of the ones that we have seen. There are no indications that the model in (1) should be altered in any significant way. In fact, the model appears to fit very well.

There are several issues, though, that require additional investigation:

1. Choice of criteria for deleting observed values and classifying them as additive outliers (i.e., those values that will not be used for prediction for nonsampled companies). The current method is based on manual review and classification of a unit as an additive outlier. Units classified as additive outliers generally were hard coded as such in the estimation program. There was no routine review of their status.

In our evaluation we have used automatic outlier procedures to classify an observation to be an outlier if the absolute value of the standardized residual exceeds 3.5, and deemed an observation to be influential if DFFITS exceeds $2/n^{1/2}$. Further assessment of this approach may lead to more consistent treatment of unusual observations in the future.

2. Composition of “estimation groups” (i.e., post-strata)

We have been considering the four and ten region groups described above. There are other alternatives based on temperature/climate and type of ownership.
More generally it would be good to have a sound method of forming estimation groups based on observed homogeneity of regression slopes.

Use of ownership as a post-stratification factor has been investigated and it appears that the regression relationships for IOUs and non-IOUs are somewhat different. While some of the components of the latter may have different slopes, the subsets are likely to be too small to be useful.

3. Use of macro-level longitudinal data

We have plotted for each variable (sales, revenue) and each end-user (residential, commercial, industrial, total) the estimated regression coefficients corresponding to each region for each of 24 consecutive months (January 2002 to December 2003). There are two plots for each combination of variable and end-user; i.e., (a) using all of the data, and (b) using all of the data except for those deemed outliers or influential observations. There are analogous plots for the residual standard deviation.

The second set of figures, labelled “Macro-level longitudinal estimates,” has twelve such plots. For each of residential, commercial and industrial sales, there are separate time plots for the estimated regression coefficients and the estimated residual standard deviations. Finally, for each these six choices there is a plot using all of the data and one using all of the data except for those deemed outliers or influential observations.

Clearly, there are seasonal patterns in the estimated regression coefficients. These could potentially be used to form empirical prior distributions which then could be used to improve the precision of estimation of the current value of the regression coefficient for a region. For some variables the estimates of the residual standard deviation are stable over time, and these values could also be used to improve the precision of estimation of the residual standard deviation on the current occasion.

4. Use of micro-level longitudinal data

We have also considered the time series behavior for individual companies within a region. For each of twenty four consecutive months (for specified variable and end-user) we have plotted for each company in a region its monthly value (normalized by its annual value/12). In this way we can look at the variation in the slopes of the individual companies.

The third set of Figures, labelled “Micro-level longitudinal data,” has two such plots, each for residential sales. One is for the NEC region, the second for the NEW region (defined earlier in this paper). Note that the estimated regression coefficients for each region are superimposed using thick blue lines.

The purpose of these plots is to assess the validity of the assumption that the seasonal patterns of the largest companies are approximately the same as the seasonal patterns for the smallest – the implicit assumption of the methodology.

5. Methodology to pool similar regression coefficients

One may obtain greater precision for inference about $\{\beta_s : s = 1,..., L\}$ in (1) by pooling appropriate data from the L regions. Standard “shrinkage” methods imply that the posterior expected value of β_s, say, has the form

$$\mathrm{E}\,(\beta_s \mid y) = \lambda_s \hat{\beta}_s + (1 - \lambda_s)\hat{\beta} \qquad (2)$$

where y denotes the observed data, $\hat{\beta}_s$ is the weighted least squares estimator of $\beta_s, \hat{\beta} = \sum \lambda_i \hat{\beta}_i / \sum \lambda_i$ and $\lambda_s = \delta^2 / \{\delta^2 + V(\hat{\beta}_s)\}$.

Here, it is assumed that, given $(\nu, \delta^2), \beta_1,..., \beta_L$ are independent, identically distributed random variables with mean ν and variance δ^2, and that ν has a locally uniform prior distribution. However, (2) does not recognize that the $\{\beta_1,..., \beta_L\}$ may not satisfy the assumption that they are independent and identically distributed random variables.

Methodology in Malec and Sedransk (1992), Evans and Sedransk (2001) and Sedransk and Yan (2004) can be used to relax this assumption. Looking at a single month in the plots of the estimated regression coefficients over time (e.g., for residential sales) it appears that the regression coefficients may be clustered, and, thus, the proposed methodology will be appropriate.

The main idea is to consider all partitions of the L regions. For example, if L=6 a few of the partitions are {(123), (456)}, {(12), (3456)}, {(13), (2456)}, and {123456}. (Note that the last partition in this list corresponds to the assumption made using the "shrinkage" methodology.) Conditional on a partition, g, it is assumed that (a)all of the regression coefficients in a subset (e.g., (123) in the first partition above) are independent and identically distributed but with an expected value specific to that partition *and* subset, and (b)there is independence over the subsets in a partition. Inference about β_s is made conditionally on partition g and then averaged over the set of partitions. That is,

$$E(\beta_s \mid y) = \sum_g E(\beta_s \mid g, y) p(g \mid y)$$

weights the conditional posterior means by the posterior probabilities associated with the L partitions. The posterior variance of β_s properly accounts for the uncertainty about the grouping of the L regression coefficients into subsets.

Questions for the Committee

1.Have you suggestions about different statistics and/or criteria for detecting outliers and influential observations?

2.Have you suggestions for forming estimation groups (post-strata) starting from (a)company level data, or (b)data from a set of geographical regions?

3.Do you have suggestions for other uses of the macro-level longitudinal data?

4.Do you have suggestions for use of the company level longitudinal data?

5.Do you have suggestions about alternative methodology for using a relatively large number of estimation groups and then using analytical tools (e.g., the "uncertain pooling" in point 5 above) for aggregation?